A National Data Grid Infrastructure for Sharing Biospecimens in Early Cancer Detection

Crichton, Daniel[*1], Kincaid, Heather[2], Johnsey, Donald[3], Kelly, Sean[1], Srivastava, Sudhir[3]
[1]NASA Jet Propulsion Laboratory, Pasadena, CA, USA; [2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA; [3]National Cancer Institute, Bethesda, MD, USA

Eradication of death and suffering from cancer by 2015 is a major goal of the National Cancer Institute (NCI) and is dependent on validation and access to data across a wide-variety of programs and projects. Capturing, accessing and correlating data that is managed by independent, distributed, heterogeneous information systems is difficult due to different semantic software models and systems that are operated by research labs across the country. The NCI's Early Detection Research Network (EDRN) comprises and integrates a large number of cancer research institutions into a research enterprise to support the discovery and validation of early detection cancer biomarkers. The diversity of both the institutions and the collaborative focus of the EDRN has created the need for establishing an infrastructure for capturing and sharing a variety of research data sets.

The EDRN has successfully developed an informatics infrastructure that enables scientists to access databases related to specimen databanks from nine institutions across the country. The Fred Hutchinson Cancer Research Center, the National Cancer Institute and NASA's Jet Propulsion Laboratory (JPL) teamed up to build the infrastructure reusing a software middleware framework that JPL originally developed to support the space science program. The middleware, named the Object Oriented Data Technology (OODT) framework, provides the ability to capture, locate, access and integrate heterogeneous data sets and data products into a virtual data grid infrastructure providing scientists a unified view of a data sets that may be geographically distributed across several computing environments. The EDRN Resource Network Exchange or ERNE has been established to enable access to a virtual database whereby scientists can search and retrieve details about specimens located and independently managed by the participating institutions. In addition, EDRN is developing a validation infrastructure as a core component of the informatics infrastructure in order to support analysis and integration of data sets captured during a study. EDRN has developed a common software and data architecture that provides not only a solution for sharing of biospecimen data sets, but a long term architecture for establishing a knowledge environment for early detection.